DOCUMENT RESUME

ED 264 288 TM 850 776

AUTHOR Doron, Rina
TITLE The "Average System" - A New Technique for Equating Scores on Different Tests Constructed from an Item Bank.
PUB DATE Dec 84
NOTE 17p.; Paper presented at the International Conference on Education in the 90's (1st, Tel-Aviv, Israel, December 16-19, 1984). Prepared at the Center for Technological Education, Holov, affiliated with Tel-Aviv University.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Arithmetic; Comparative Analysis; *Equated Scores; Grade 5; Intermediate Grades; *Item Banks; *Measurement Techniques; Reliability; Scaling; *Test Construction; Test Validity
IDENTIFIERS *Rasch Model; *Rasch Scaled Scores

ABSTRACT

An alternative procedure is examined to compare test scores which does not require the utilization of computers and therefore, can be easily employed by classroom teachers. This technique is called the Average System: a simple way to score examinees using information obtained from a representative sample (or samples). Examinees are ranked on a four point ordinal scale, using information on item difficulty derived from different groups in the representative sample, for estimating average scores of different groups in the population. This system enables comparison of examinees' scores on different tests, constructed out of an item bank. An examinee's score is related to the average appropriate to the test, and an ordinal score assigned. Ordinal scores for different examinees can then be compared without regard to the test used. The validation of the Average System is discussed. Information about item difficulty according to the Rasch model was also derived to permit the validation of the Average System. The criteria for validation is presented and the method to test for item consistency and matching are described. Results indicate that the Rasch Model and the Average System are interchangeable regarding examinee ranking and measurement temporal stability. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED264288

CENTER FOR TECHNOLOGICAL EDUCATION HOLON

AFFILIATED WITH TEL AVIV UNIVERSITY

THE "AVERAGE SYSTEM" - A NEW TECHNIQUE FOR EQUATING SCORES ON DIFFERENT TESTS CONSTRUCTED FROM AN ITEM BANK

By

RINA DORON, Ph.D.

THE HOLON CAMPUS, 52 GOLOMB ST., P.O.B. 305, HOLON, ISRAEL.

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Doron

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PAPER PRESENTED AT

THE 1ST INTERNATIONAL CONFERENCE

ON EDUCATION IN THE 90'S

TEL-AVIV, DECEMBER 16-19, 1984

TM 850 776

ACKNOWLEDGEMENTS

The author wishes to thank Professor Arieh Levy and Professor Norman Rudy for their support in preparing this paper.

In the field of achievement testing one frequently encounters the problem of equating scores which have been derived from different tests. This may occur while conducting longitudinal studies. One can hardly present the same configuration of items for a pupil at a low grade and several years later at a higher grade level. The content of instruction changes year after year and the test items have to reflect changing contents of instruction.

Similarly, while conducting a large scale achievement survey in a hetergeneous population, it may be inappropriate to administer the same test to groups of students whose content of studies differed from each other. This may be the case either as a result of differential school programs for various ability groups or as a result of differential content emphasis reflecting different approaches to teaching a given subject.

These are only a few examples of the case of different tests within the framework of a single study, while the researcher is interested in assigning scores to these tests on a common scale which enables comparisons to be made. For dealing with such situations, probably the best known method is the technique of sample-free calibration developed by Rasch, 1960, 1966,(a) (b); Choppin, 1968, 1969, 1976, 1978 (a) (b)..

This method can be used whenever a small set of identical items appears in two or more tests, the scores of which have to be compared. In principle, the Rasch approach provides a satisfactory solution to this vexing problem. Nevertheless, in practice, when such comparisons have to be performed by teachers at classroom or at school level, without easy access to computers, one encounters difficulties in dealing with scores derived from these individual tests.

It should be mentioned that Wright developed a technique for establishing scores without the use of a computer (Wright, 1968; Wright & Douglas 1975; Wright & Stone, 1978). This technique, however, demands that test creators operate under extremely rigid conditions. It is very difficult to construct tests when, on the one hand, Wright's requirements must be met, while, on the other, the criteria of educational objectives must be fulfilled.

The purpose of this study is to examine an alternative procedure to compare test scores which does not require the utilization of computers and therefore, can be easily employed by classroom teachers. This technique is called the Average System.

While the Rasch model is based on computation procedures which can be employed whenever two different tests have a set of identical anchor items, without regard to whether the two different tests were administered to samples of equal ability level, the Average System requires a more restrictive setting for deriving basic item parameters which enable comparison of such scores. It is necessary that the different items of the tests to be compared should be first administered to a small representative sample of pupils from which the basic parameters of the items should be derived. While this restriction severely limits the calibration possibilities of the suggested Average System by comparison with the fully sample-free calibration possibilities of the Rash model, in practice teachers can more easily overcome such restrictions than those imposed by the need to use large computers and computer programs. Thus, for example, if within a single school one would like to administer various tests to children in grade 6, a representative group of 6th grade students can be formed who will respond to all the items in all the tests, or several groups of 6th grade students of representative ability can be employed, with each group

responding to all the items on a single test.

THE AVERAGE SYSTEM

This is a simple way to score examinees using information obtained from a representative sample (or samples). The information on item difficulty includes measures obtained from the total sample and separately from the "low" achievers and the "high" achievers of the sample.

It is known that a test average is actually a total of the average of the items which compose it:

$$\bar{X} = \sum_{i=1}^{k} P_i$$

where $\bar{X}$ = test average

P_i = difficulty of item i

k = number of items on the test

Therefore, the test average for any population may easily be estimated from the average of the items of which that test is composed. When information on an item's difficulty (p_i) is given for the entire sample (p), as well as for groups which are termed "high"(P_H) and "low", (P_L) then, one may estimate both the average of the entire sample, $\bar{X} = \Sigma P$, as well as the average of the "highs", $\bar{X}_H = \Sigma P_H$ and of the "lows", $\bar{X}_L = \Sigma P_L$ for every combination of k items.

In this way, four levels of achievement may be established, and examinees can be classified in accordance with their actual achievement level on that test.

Below $\bar{X}_L$ - Below the "low" average - 1

($\bar{X}_L$, $\bar{X}$) - Between the "low" average and the sample average - 2

($\bar{X}$, $\bar{X}_H$) - Between the sample average and the "high" average - 3

Above $\bar{X}_H$ - Above the "high" average - 4

This allows the ranking of examinees on a four point ordinal scale, using information on item difficulty derived from different groups in the representative sample, for estimating average scores of different groups in the population.

The Average System enables comparison of examinees' scores on different tests, constructed out of an item bank. Suppose, for example, that Test A is composed of m items of the bank and test B of n items. Using the established item difficulties (available for each of the three groups in the representative sample), the three average scores are developed for test A and test B, separately. An examinee's score is then related to the average appropriate to his test, and an ordinal score assigned. Ordinal scores for different examinees can then be compared without regard to the test used.

For example, both an examinee whose ordinal score is 4 on test A, and an examinee whose ordinal score is 4 on test B, have achievements which are above the average of the group defined as "high" in the representative sample.

THE VALIDATION OF THE AVERAGE SYSTEM

The Construction of the Item Bank and the derivation of Parameters

To validate the Average System as a satisfactory approach for equating scores from various tests, an experiment was carried out. A 230 item collection dealing with the content of fifth grade arithmetic was developed, and 23 sets of 10 item tests were constructed from it. The 23 sets were administered to 15 different samples having comparable abilities (i.e. having similar frequency distributions of mathematical ability). The tests were distributed to the groups in such a way that 21 out of the 23 tests were used as anchor items. Following this, the parameters P_L, P and P_H were calculated for each item.

In other words, information was derived on the difficulty of each and every item for groups which were defined as "high" and "low" achievers, as well as for the entire sample. Information about item difficulty according to the Rasch model was also derived in order to permit the validation of the Average System.

THE CRITERIA FOR VALIDATION

The criteria for validating the Average System were as follows:

1. The correlations between scores on item-bank based tests computed according to the Rasch model, and scores computed according to the Average System on the same tests, (the concurrent validity of the Average System as the newer of the two), were calculated.
2. Comparison was made of the two techniques with regard to the temporal stability of scores given on tests constructed from the item bank. In other words, if the same group of examinees is tested twice with a lapse of time between tests which use different compositions of items, how similar will be the two scores using the Average System, as compared to the similarity using the Rasch model? The temporal stability of test scores is a well-known indicator of reliability. However, this reliability is not only a product of the method of a score's computation, but also of other contributing factors. Two terms are suggested in this paper to describe temporal stability of scores: "consistency" and "matching".

   Item bank consistency - An item bank may be considered "consistent" if the same examinees maintain the same ranks when tested on different tests (combinations of items) -- whose items have all been chosen from that same item bank.

Item bank matching - An item bank may be considered to "match" if the examinees obtain the same scores on two different tests. (This requirement is very extreme, of course). Obviously, when there is "Matching" there will always be "consistency" (but not vice versa).

Both techniques are based on examinees' "Raw scores", computed according to the number of questions he answers correctly, without regard to the particular combination of items which are answered correctly. If we were not seeking normative comparison with other groups in the population, then "Raw scores" would suffice.

These are scores whose temporal stability depends on the reliability of the test itself. The question we raised, therefore, was: How much temporal stability, are we losing or gaining by using the Rasch model or the Average System, as compared with the temporal stability we would gain were we using only "Raw Scores"?

METHOD

The following steps were taken to test for consistency and matching:

a. A comparison of examinee achievement on a test, with the achievement of this same group on a different test constructed from the same bank and administered several days later.

It was decided to administer the tests to two groups of pupils:

LO - Pupils in schools of predominantly under-privileged familes.

Hi - Pupils in schools of predominantly privileged families.

Score Consistency was tested for by reference to the coefficients of correlation between scores in the two test administrations.

Score matching was tested for by reference to the average of differences between scores and to the standard deviation of the difference between scores in the two test administrations.

Given that we may assume full matching between two series of scores if both their average difference and the standard deviation of the indifferences are very close to zero.

Due to the fact that the scores in the two systems are on different scales, it was decided to translate the scores in both test systems to standard scores (Z) and to compare the average differences of the standard score ($|\overline{d}|_z$) and the standard deviation of the difference of the standard scores (Sd_z).

As for the concurrent validity of the Average System, coefficients of correlation were calculated for examinees' scores as established by the Average System and scores established by the Rasch model, on each of the two administrations.

RESULTS

Table 1 exhibits the correlation coefficents relating to the concurrent validity of the Average System.

It reveals that the coefficients of correlation between scores determined using the Rasch model and scores determined using the Average System are high (between .76 to .92). These findings lead to the conclusion that there is no substantial difference between the two systems with regard to examinee ranking (using the same test). These data support the conclusion that the Average System possesses concurrent validity.

The data regarding the temporal stabilities of the Average System and the Rasch Model are given in Tables 2 & 3. The findings in Table 2 show that the coefficents of correlation between "Raw scores" on the first and second administrations were higher than those established using the Rasch model

and the Average System. However, none of the differences were statistically significant. That is to say, when scores are determined using the Rasch model and the Average system there is a certain decrease in the consistency of these scores as compared with the "Raw scores", but this decrease is not significant. Table 2 shows also that within the Lo group the correlation between scores established by the Rasch model on the first and the second administrations is lower than that established using the Average System,within the Hi group the correlation between scores established by the Rasch model was higher than that established by the Average System. But, in no instance were any of the differences statistically significant, which means that there is no meaningful difference between the two techniques in so far as their consistency is concerned.

As can be seen in the results presented in Table 3, the means of the standardized scores' differences ($|\bar{d}|_z$). of the "Raw Scores" are higher than those found for the Rasch Model and the Average System. However, the standard deviation of these differences (Sd_z) show a reverse trend, i.e. they are the lowest for the "Raw Scores". This was found both in the case of the Hi grop and the Lo group.

It can also be seen in the Lo group while the $|\bar{d}|_z$ of the Rasch Model was lower than the $|\bar{d}|_z$ of the Average System, the Sd_z of the Average System was lower than the Sd_z of the Rasch Model. In the Hi group, the $|\bar{d}|_z$ of the two techniques were found to be zero, and the Sd_z of the Rasch Model was found to be somewhat lower than that of the Average System. According to these findings,it can be argued that in so far as the temporal stability of the score matching is concerned, neither one of the two techniques can be

said to be preferable over the other.

CONCLUSION

All of the above leads to the conclusion that the Rasch Model and the Average System are interchangeable with regard to examinee ranking, as well as with regard to measurement temporal stability.

Therefore, when teachers have no access to computers and lack the mathematical background needed for using the Rasch model, the simple technique of the Average System could be used, since the latter technique is relatively easier to explain and employ.

TABLE 1

Correlations between scores as established by the Rasch Model & the Average System on the 1st and 2nd administrations (determined for Lo & Hi pupils).

Group level	N	1st admin. r	2nd admin. r
Lo	16	.90	.92
Hi	29	.90	.76

TABLE 2

Correlations between scores on]st and 2nd administrations as established by the Rasch Model, by the Average System and the "Raw Scores". (dtermined for Lo & Hi).

Group level	N	Rasch-Model r	Average System r	"Raw Scores" r
Lo	16	.87	.89	.92
Hi	29	.73	.62	.74

TABLE 3

Means and standard deviations of standardized scores' differences between 1st & 2nd administrations of: Scores established by the Rasch Model, by the Average System and the "Raw Scores" (determined for Lo & Hi pupils).

Group Level	N	Rasch-Model $\lvert\bar{d}\rvert_z$	Sd_z	Average - System $\lvert\bar{d}\rvert_z$	Sd_z	"Raw Scores" $\lvert\bar{d}\rvert_z$	Sd_z
Lo	16	.01	.65	.17	.50	.30	.25
Hi	29	.01	.60	.00	.82	.54	.47

REFERENCES

Choppin, B. An item bank using sample-free calibration, Nature,1968, 219, 872-970.

Choppin, B. An item bank using sample-free calibration. In: Wood R.L., Skwinik, L.S. Item banking a method for proceeding school based examinations and nationally comparable grades. National Foundation for Educational Research in England & Wales, 1969.

Choppin, B. Recent developments in item banking. In: D.N. Le Gruitjer & L.J. Vanderkamp (eds.), Advances in psychological and educational measurement, London, Wiley, 1976.

Choppin, B. Item Banking and the Monitoring of Achievement. Research in Progress No. 1 Slough, NFER, 1978 (a).

Choppin, B. (Ed), Psychometric developments relating to item banking, a selected annotated bibliography. The NFER 1978. (b).

Rasch, G. Probabilistic models for some intelligence and attainment tests.
Copenhagen: Danish Institute for Educational Research, 1960, Chapters V-VII, X.

Rasch, G. Individualistic approach to item analysis in Reading in Mathematical Social Science. Edited by Lazarsfield and Henry. Chicago: Science Research Associates Inc. 1966, 89-107 (a).

Rasch, G. An item analysis which takes individual differences into account British Journal of Mathematical and Statistical Psychology, 1966, 19 49-57 (6)

Wright, B. Sample-free test calibration and person measurements. In proceedings of the 1967 Invitational Conference on testing problems. Princeton, N.J.: Educational testing Service, 1968, 85-101.

Wright, B.D. & Douglas, G.A. Best test design and self-tailored testing. Research Memorandum No. 19, Statistical laboratory, Department of Education, University of Chicago, 1975.

Wright, B.D., & Stone M.A. Best test design: A handbook for Rasch Measurement Palo Alto: Scientific Press, 1978.